The Green500 List: Encouraging Sustainable Supercomputing

Wu-chun Feng and Kirk W. Cameron
Virginia Tech

The performance-at-any-cost design mentality ignores supercomputers' excessive power consumption and need for heat dissipation and will ultimately limit their performance. Without fundamental change in the design of supercomputing systems, the performance advances common over the past two decades won't continue.

Although there's now been a 10,000-fold increase since 1992 in the performance of supercomputers running parallel scientific applications, performance per watt has only improved 300-fold and performance per square foot only 65-fold. In response to the lagging power and space-efficiency improvements, researchers have had to design and construct new machine rooms, and in some cases, entirely new buildings.

Compute nodes' exponentially increasing power requirements are a primary driver behind this less efficient use of power and space. In fact, the top supercomputers' peak power consumption has been on the rise over the past 15 years, as Figure 1 shows.

Today, the 10 most powerful supercomputers on the TOP500 List (www.top500.org) each require up to 10 megawatts of peak power—enough to sustain a city of 40,000. And even though IBM BlueGene/L, the world's fastest machine, was custom-built with low-power components, the system still consumes several megawatts of power. At anywhere from $200,000 to $1.2 million per megawatt, per year, these are hardly low-cost machines.

THE ENERGY CRISIS IN SUPERCOMPUTING

Power is a disruptive technology that requires us to rethink supercomputer design. As a supercomputer's nodes consume and dissipate more power, they must be spaced out and aggressively cooled. Without exotic cooling facilities, overheating makes traditional supercomputers too unreliable for application scientists to use. Unfortunately, building exotic cooling facilities can cost as much as the supercomputer itself, and operating and maintaining the facilities costs even more.

As "The Energy-Efficient Green Destiny" sidebar details, the low-power supercomputer that we developed was extremely reliable, with no unscheduled downtime in its two-year lifespan, despite residing in a dusty warehouse without cooling, humidification, or air filtration. The hourly cost of such downtime ranges from $90,000 for a catalog sales operation to nearly $6.5 million for a brokerage operation, according to Contingency Planning Research's 2001 cost-of-downtime survey.

There's still no guarantee that the supercomputer won't fail, as Table 1 illustrates. Total cost of ownership now exceeds initial acquisition costs.

Performance at any cost

The performance-at-any-cost supercomputer design paradigm is no longer feasible. Clearly, without significant change in design, the performance gains of the past two decades won't continue. Unfortunately, performance-only metrics don't capture improvements in power efficiency. Nonetheless, performance-only metrics derived from the Linpack benchmarks and Standard Performance Evaluation Corp.'s (SPEC) code suite have significantly influenced the design of modern high-performance systems, including servers and supercomputers.

Published by the IEEE Computer Society
0018-9162/07/$25.00 © 2007 IEEE

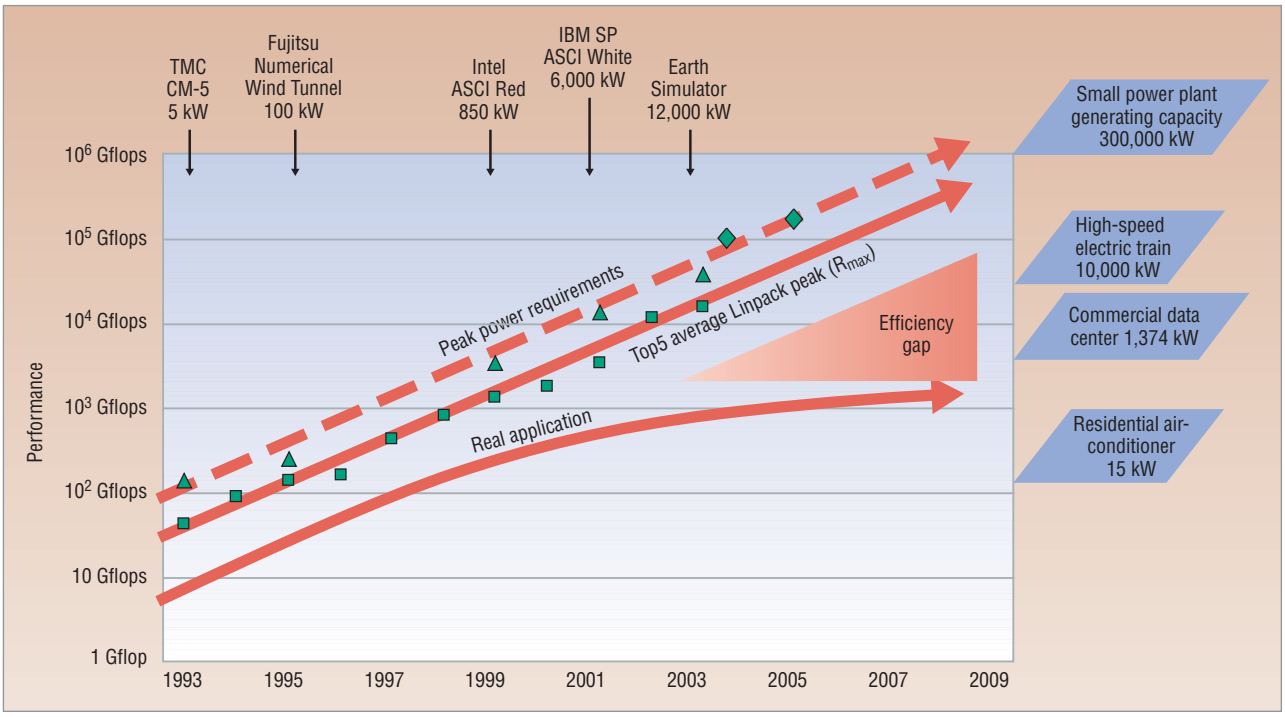

Figure 1. Rising power requirements. Peak power consumption of the top supercomputers has steadily increased over the past 15 years.

The Energy-Efficient Green Destiny

As a first step toward reliable and available energy-efficient supercomputing, in 2002 we built a low-power supercomputer at Los Alamos National Laboratory. Dubbed Green Destiny, the 240-processor supercomputer took up 5 square feet (the size of a standard computer rack) and had a 3.2-kilowatt power budget (the equivalent of two hairdryers) when booted diskless.[1,2] Its 101-gigaflop Linpack rating (equivalent to a 256-processor SGI Origin 2000 supercomputer or a Cray T3D MC1024-8) would have placed it at no. 393 of the 2002 TOP500 List.

Garnering widespread media attention, Green Destiny delivered reliable supercomputing with no unscheduled downtime in its two-year lifetime. It endured sitting in a dusty warehouse at temperatures of 85-90 degrees Fahrenheit (29-32 degrees Celsius) and an altitude of 7,400 feet (2,256 meters). Furthermore, it did so without air-conditioning, humidification control, air filtration, or ventilation.

Yet despite Green Destiny's accomplishments, not everyone was convinced of its potential. Comments ranged from Green Destiny being so lower power that it ran just as fast when it was unplugged to the notion that no one in HPC would ever care about power and cooling.

However, in the past year, we've seen a dramatic attitude shift with respect to power and energy, particularly in light of how quickly supercomputers' thermal power envelopes have increased in size, thus adversely impacting the systems' power and cooling costs, reliability, and availability.

The laboratory's Biosciences Division bought a Green Destiny replica about six months after Green Destiny's debut. In 2006, we donated Green Destiny to the division so it could run a parallel bioinformatics code called mpiBLAST. Both clusters are run in the same environment, yet half of the nodes are inoperable on the replica, which uses higher-powered processors. Hence, although the original Green Destiny was 0.150 gigahertz slower in clock speed, its productivity in answers per month was much better than the faster but often inoperable replica.

Green Destiny is no longer used for computing, and resides in the Computer History Museum in Mountain View, California.

References

1. W. Feng, "Making a Case for Efficient Supercomputing," *ACM Queue*, Oct. 2003, pp. 54-64.
2. W. Feng, "The Importance of Being Low Power in High-Performance Computing," *Cyberinfrastructure Technology Watch*, Aug. 2005, pp. 12-21.

Table 1. Reliability and availability of large-scale computing systems.

System	Processors	Reliability and availability
ASC Q	8,192	Mean time between interrupts: 6.5 hours, 114 unplanned outages/month Outage sources: storage, CPU, memory
ASC White	8,192	Mean time between failures: 5 hours (2001) and 40 hours (2003) Outage sources: storage, CPU, third-party hardware
PSC Lemieux	3,016	Mean time between interrupts: 9.7 hours Availability: 98.33 percent
Google (estimate)	450,000	600 reboots/day; 2-3 percent replacement/year Outage sources: Storage and memory Availability: ~100 percent

Source: D.A. Reed

Developing new metrics

Performance-only metrics are likely to remain valuable for comparing existing systems prior to acquisition and helping drive system design. Nonetheless, we need new metrics that capture design differences in energy efficiency. For example, two hypothetical high-performance machines could both achieve 100 teraflops running Linpack and secure a high equivalent ranking on the TOP500 List. But enable smart-power-management hardware or software[1,2] on one machine that can sustain performance and reduce energy consumption by 10 percent, and the TOP500 rankings remain the same.

Unfortunately, metric development is fraught with technical and political challenges. On the technical side, operators must perceive the metric and its associated benchmarks as representative of the workloads typically running on the production system. On the political side, metrics and benchmarks need strong community buy-in.

THE GREEN500 LIST

We've been working to improve awareness of energy-efficient supercomputer (and data-center) design since the turn of the century. After interaction with government agencies, vendors, and academics, we identified a need for metrics to fairly evaluate large systems that run scientific production codes. We considered a number of methodologies for use in ranking supercomputer efficiency. To promote community buy-in, the initial Green500 List used a single metric and widely-accepted workload while the intent is to extend the Green500 methodology to eventually include rankings for a suite of parallel scientific applications.

The Green500 List ranks supercomputers based on the amount of power needed to complete a fixed amount of work. This effort is focused on data-center-sized deployments used primarily for scientific production codes. In contrast, the SPECPower subcommittee of SPEC is developing power-performance efficiency benchmarks for servers running commercial production codes. The diverse types of evaluations that efforts like the Green500 and SPECPower (www.spec.org/ specpower) provide will give users more choice in determining efficiency metrics for their systems and applications.

Measuring efficiency

In the Green500 effort, we treat both performance (speed) and power consumption as first-class design constraints for supercomputer deployments.

Speed and workload. The supercomputing community already accepts the flops metric for the Linpack benchmark, which the TOP500 List uses. Although TOP500 principals acknowledge that Linpack isn't the be-all or end-all benchmark for high-performance computing (HPC), it continues to prevail despite the emergence of other benchmarks. As other benchmark suites gain acceptance, most notably the SPEChpc[3] and HPC Challenge benchmarks,[4] we plan to extend our Green500 List methodology as mentioned. For now, since the HPC community seems to identify with the notion of a clearly articulated and easily understood single number that indicates a machine's prowess, we opt to use floating-point operations per second (flops) as a speed metric for supercomputer performance and the Linpack benchmark as a scalable workload.

EDn metric. There are many possibilities for performance-efficiency metrics, including circuit design's EDn metric—with E standing for the energy a system uses while running a benchmark, D for the time to complete that same benchmark,[5-8] and n a weight for the delay term. However, the EDn metrics are biased when applied to supercomputers, particularly as n increases. For example, with large values for n, the delay term dominates so that very small changes in execution time impact the metric dramatically and render changes in E undetectable in comparisons.

Flops per watt. For the Green500 List, we opted to use flops per watt for power efficiency. However, this metric might be biased toward smaller supercomputing systems. A supercomputer's wattage will scale (at least) linearly with the number of compute nodes while the flops performance will scale (at most) linearly for embarrassingly parallel problems and sublinearly for all other problems. This implies smaller systems would have better ratings on such a scale.

Nonetheless, flops per watt is easy to measure and has traction in the scientific community. Furthermore, we can reduce the bias toward small systems by ranking systems that first achieve a minimum performance rating. We simply set a minimum flops threshold for entry into the Green500 List and allow bigger super-

computers to rerun their Linpack benchmark, if desired, to meet this minimum threshold and obtain the corresponding power consumption during a benchmark's rerun. That is, the Green500 List ranks supercomputers based on the amount of power needed to complete a fixed amount of work at a rate greater than or equal to the minimum flops threshold.

Measuring power consumption

Even after choosing the benchmark and power-efficiency metric, issues surrounding the selection of the flops-per-watt metric for a given supercomputer remained unresolved. First, starting with the metric's numerator, what will be the minimum flops threshold for entry into the Green500 List? We intend to use the flops rating that the no. 500 supercomputer achieves on the latest TOP500 List as the Green500 List's minimum flops threshold.

With the numerator addressed, this leaves wattage as the denominator. Surprisingly to some, the denominator for flops per watt might be more difficult to determine, since there are many permutations for what we could measure and report. For example, we could

- measure the entire supercomputer's power consumption,
- measure a single supercomputer node and extrapolate it to the entire supercomputer, or
- use the manufacturers' advertised peak power numbers (as we used in Figure 1).

Measuring wattage for a supercomputer the size of a basketball court is difficult. However, using advertised peak power numbers could result in over-inflation of the power numbers. We suggest measuring a single compute node's power consumption and multiplying by the number of compute nodes (loosely defining a node as an encased chassis, whether the chassis has the form factor of a standard 1U server or an entire rack).

Power meters. To measure power consumption, we propose using a power meter that can sample consumption at granularities of one second or less. The digital meters range in capability from the commodity Watts Up? Pro (www.wattsupmeters.com) to the industrial-strength Yokogawa WT210/WT230 (http://yokogawa.com/tm/wtpz/wt210/tm-wt210_01.htm). Figure 2 shows a high-level diagram of how a digital power meter measures a given system under test (single compute node) via a common power strip and logs the measurements to a profiling computer.

Duration. We also need to address how long to measure power and what we should record and report. Given the meters' recording capabilities, we suggest measuring and recording power consumption for the duration of the Linpack run and using the average power consumption over the entire run. Coupling average power consumption with the Linpack run's execution time adds inferred overall energy consumption, where energy is average power multiplied by time.

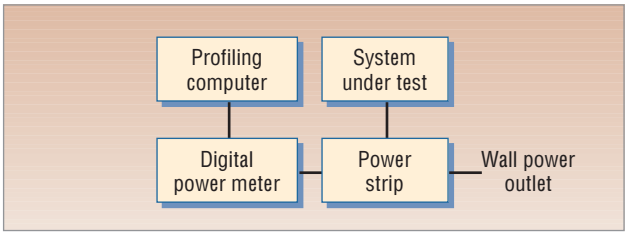

Figure 2. Power-measurement infrastructure. A digital power meter measures a system under test via a common power strip and logs the measurements to a profiling computer.

Cooling. Finally, we considered whether to include cooling-facility power consumption in the measurement. We decided against inclusion because the Green500 List is intended to measure the supercomputer's power efficiency, rather than cooling systems (which vary widely in power efficiency). Even if we considered cooling for the supercomputer under test, it would be difficult to break out and measure cooling's specific contribution for one supercomputer, given that cooling facilities are designed to support all the machines in a given machine room.

TOP500 VERSUS GREEN500

Table 2 presents the Green500 and TOP500 rankings of eight supercomputers, as well as their flops ratings and their peak power usage. This list also shows the results of using the flops-per-watt metric for these supercomputers using their peak performance number (for peak power efficiency) and their Linpack performance number (for actual power efficiency).

As mentioned, using peak power numbers for comparisons isn't optimal. Nonetheless, the relative comparisons using peak power numbers are useful to gauge power-efficiency progress. Beginning with the November 2007 Green500 List, we'll use metered measurements in rankings whenever available. As the list matures, we anticipate metering and verifying all measurements.

Various presentations, Web sites, and magazine and newspaper articles provide the source for these peak power numbers. For the IBM BlueGene/L supercomputer at Lawrence Livermore National Laboratory, the TOP500 wiki reports 1.5 MW as its peak power consumption. LLNL's Web site reports that 7.5 MW is needed to power and cool ASC Purple, while Eurekalert estimates it uses 8 MW.

According to LLNL, for every watt of power the system consumes, 0.7 watts of power is required to cool it. Hence, the power required to merely run ASC Purple would be between 4.4 and 4.7 MW, which matches the 4.5 MW number provided in a presentation at a BlueGene/L workshop.

Jaguar at Oak Ridge National Laboratory is a hybrid system consisting of 56 XT3 cabinets and 68

Table 2. June 2007 Green500 and TOP500 rankings.

Green500 rank (power efficiency)	Supercomputer	Peak performance (Gflops)	Linpack performance (Gflops)	Peak power (kW)	Peak power efficiency (Mflops/W)	Green500 rank (peak power efficiency	Power efficiency (Mflops/W)	TOP500 rank
1	BlueGene/L	367,000	280,600	1,200	305.83	1	233.83	1
2	MareNostrum	94,208	62,630	1,344	70.10	2	46.60	9
3	Jaguar	119,350	101,700	2,087	57.19	4	48.73	2
4	System X	20,240	12,250	310	65.29	3	39.52	71
5	Columbia	60,960	51,870	2,000	30.48	5	25.94	13
6	ASC Purple	92,781	75,760	4,500	20.62	6	16.84	6
7	ASC Q	20,480	13,880	2,000	10.24	7	6.94	62
8	Earth Simulator	40,960	35,860	7,680	5.33	8	4.67	20

Table adapted from a figure provided by NXP Semiconductors.

XT4 cabinets. The peak power consumption of an XT3 cabinet is 14.5 kW while the XT4 cabinet is 18.75 kW, as per Cray datasheets. Thus, the aggregate peak power of Jaguar is about 2 MW.

The 4,800-processor MareNostrum debuted fifth on the June 2005 TOP500 List with an estimated 630 kW power budget to run the machine. More recently, Barcelona Supercomputing Center's MareNostrum was upgraded and expanded into a 10,240-processor BladeCenter JS21 Cluster. If we extrapolate from the original MareNostrum's 630-kW power budget, the 10,240-processor MareNostrum would have a power budget of 1.3 MW.

For the Columbia supercomputer at NASA Ames Research Center, the reported power usage just to run the system is 2 MW. The thermal design power of Itanium-2 processors is 130 watts, so it takes 1.33 MW to run the 10,240 processors in the Columbia. Therefore, 2 MW seems reasonable if Columbia's other components use only 700 kW of power, consistent with our Itanium-based server's power profile.

In November 2005, it took an estimated 1.33 MW to power Jaguar's 5,200 processors. Jaguar's doubling in size to 10,424 processors a year later raised the extrapolated power budget to 2.66 MW.

Powering and cooling Japan's 5,120-processor Earth Simulator requires 11.9 MW, enough to power a city of 40,000 and a 27,000-student university. The Earth Simulator configures the 5,120 processors into 640 eight-way nodes, where each eight-way node uses 20 kilovolt-amperes. Assuming a typical power-factor conversion of 0.6, each node then consumes 20 kVA × 0.6 = 12 kW. Thus, power consumption for the entire 640-node Simulator is 640 × 12 kW = 7,680 kW, leaving 4,220 kW for cooling.

The power budgets for ASC Q and ASC White run at approximately 2 MW, while System X at Virginia Tech consumes a paltry 310 kW, as measured directly from System X's power distribution units. As Table 2 shows, despite its large size, BlueGene/L is the only custom low-power supercomputer among the Top500. It's routinely the highest-ranking supercomputer on both the TOP500 and Green500 lists, with a performance-power ratio that's up to two orders of magnitude better than the other supercomputers in Table 2.

The power efficiencies of MareNostrum (semicommodity) and System X (commodity) are 2.5 times better than the other supercomputers, and this ranked them second and fourth on the June 2007 Green500 List, as shown in Table 2. Interestingly, Apple, IBM, and Motorola's commodity PowerPC processor drives both of these power-efficient supercomputers. On the other hand, ASC Purple, which ranked sixth on that TOP500 list, is also based on the PowerPC processor, albeit the Power5, its higher-powered relative. Power5 ultimately contributes to ASC Purple's lower power efficiency and its sixth-place ranking on the 2007 Green500.

OPERATIONAL COSTS AND RELIABILITY

Power consumption has become an increasingly important issue in HPC. Ignoring power consumption as a design constraint results in a HPC system with high operational costs and diminished reliability, which often translates into lost productivity.

With respect to high operational costs, ASC Purple has a 7.5-MW appetite (approximately 4.5 MW to power the system and 3 MW for cooling). With a utility rate of 12 cents per kW/hour, the annual electric bill for this system would run nearly $8 million. If we scaled this architecture to a petaflop machine, powering up and cooling down the machine would require approximately 75 MW. The system's annual power bill could run to $80 million, assuming energy costs remained the same.

Table 1 shows that the reliability and availability of large-scale systems, ranging from supercomputers to a large-scale server farm, is often measured in hours. Further scaling of such supercomputers and data cen-

ters would result in several failures per minute.[9] This diminished reliability results in millions of dollars per hour in lost productivity.

In light of the above, the HPC community could use an EnergyGuide sticker, such as the Green Destiny sticker shown in Figure 4. The community could also use studies showing that annual server power and cooling costs are approaching annual spending on new machines.

The HPC community needs a Green500 List to rank supercomputers on speed and power requirements, and supplement the TOP500 List. Vendors and system architects worldwide take substantial pride and invest tremendous effort toward making the biannual TOP500 List. We anticipate that the Green500 List effort will do the same and encourage the HPC community and operators of Internet data centers to design more power-efficient supercomputers and large-scale data centers. For the latest Green500 List, visit www.green500.org.

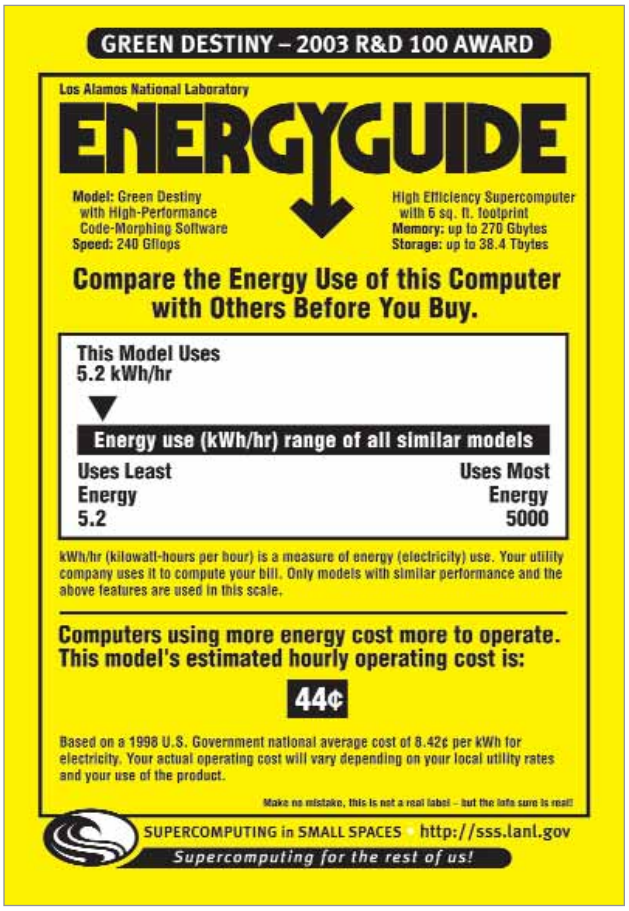

Figure 4. EnergyGuide sticker for Green Destiny. Such a sticker could remind those in the HPC community of a computer's energy use and hourly operating costs.

Acknowledgments

We thank David Bailey, Jack Dongarra, John Shalf, and Horst Simon for their support. Intel, IBM, Virginia Tech, the US Department of Energy, and the NSF (CNS 0720750 and 0615276; CCF 0709025 and 0634165) sponsored portions of this work. We also acknowledge colleagues who suggested a Green500 List after an April 2005 keynote at the IEEE International Parallel & Distributed Processing Symposium and a follow-up talk at North Carolina State University. This article is dedicated to those who lost their lives in the 16 April 2007 tragedy at Virginia Tech.

References

1. C. Hsu and W. Feng, "A Power-Aware Run-Time System for High-Performance Computing," *Proc. ACM/IEEE SC Conf.* (SC|05), IEEE CS Press, 2005, p. 1.
2. K.W. Cameron et al., "High-Performance, Power-Aware Distributed Computing for Scientific Applications," *Computer*, Nov. 2005, pp. 40-47.
3. M. Mueller, "Overview of SPEC HPC Benchmarks," BOF presentation, ACM/IEEE SC Conf. (SC|06), 2006.
4. J. Dongarra and P. Luszczek, *Introduction to the HPC Challenge Benchmark Suite*, tech. report, Univ. Tennessee, 2004; www.cs.utk.edu/~luszczek/pubs/hpcc-challenge-intro.pdf.
5. A. Martin, "Towards an Energy Complexity of Computation," *Information Processing Letters*, vol. 77, no. 2-4, 2001, pp. 181-187.
6. D. Brooks and M. Martonosi, "Dynamically Exploiting Narrow Width Operands to Improve Processor Power and Performance," *Proc. 5th Int'l Symp. High-Performance Computer Architecture*, IEEE CS Press, 1999, p. 13.
7. R. Gonzalez and M. Horowitz, "Energy Dissipation in General-Purpose Microprocessors." *IEEE J. Solid-State Circuits*, Sept. 1996, pp. 1277-1284.
8. A. Martin, M. Nystrm, and P. Penzes, *ET2: A Metric for Time and Energy Efficiency of Computation*, Kluwer Academic Publishers, 2002.
9. S. Graham, M. Snir, and C. Patterson, eds., *Getting Up to Speed: The Future of Supercomputing*, Nat'l Academies Press, 2005.

Wu-chun Feng is an associate professor of computer science and electrical and computer engineering at Virginia Tech. His research interests are high-performance networking and computing. Feng received a PhD in computer science from the University of Illinois at Urbana-Champaign. He is a senior member of the IEEE Computer Society. Contact him at feng@cs.vt.edu.

Kirk W. Cameron is an associate professor of computer science at Virginia Tech and director of its Scalable Performance Lab. His research interests are power and performance in high-performance applications and systems. Cameron received a PhD in computer science from Louisiana State University. He is a member of the IEEE Computer Society. Contact him at cameron@cs.vt.edu.

Programmable self-assembly offers an opportunity to perform computation during the fabrication process itself.